Linguistic Research 28(1), 239-256

Choosing a Reference Corpus for Keyword Calculation*

Gwang-Yoon Goh
(Yonsei University)

Goh, Gwang-Yoon. 2011. Choosing a Reference Corpus for Keyword Calculation. *Linguistic Research* 28(1), 239-256. Keywords, which are known to provide a useful way to characterize a text, are usually calculated using two word lists, one from the study corpus (SC) and the other from the reference corpus (RC). Although this notion of keywords has attracted great attention and been employed in many corpus-based language studies, the issue of what constitutes a good or appropriate RC has been left largely untouched, although an RC is generally expected to be larger in size than the SC. This paper looks into how different factors associated with the RC affect the outcome of the keyword calculation of a given SC. The results indicate that genre and diachrony are more important factors to consider than other factors when choosing an RC, especially in that the differences in these two factors, unlike those in other factors such as corpus size and varietal difference, bring about a statistically significant difference in the number of the keywords. Despite the possible effects that the size and composition of the RCs can have on keyword calculation and resulting differences in keyword results, however, keyword analysis is very robust and keywords can be plausible indicators of aboutness, regardless of the RC one chooses. Thus, the aboutness of a text should be interpreted with its possible diversity caused by the use of different RCs in mind. **(Yonsei University)**

Key Words reference corpus, study corpus, keyword, word list, corpus size, genre, national variety, diachrony

1. Introduction

A keyword normally indicates a significant word from a title or document used as an index to content. In corpus-based linguistic studies, however, the notion is defined as a word "whose frequency is unusually high in comparison with some norm" (Scott 2008: 135). In particular, although they may not be the most important words in the given text or corpus (mainly because their importance or keyness is

* I am grateful to the anonymous referees for their helpful comments and suggestions.

determined purely statistically), keywords often provide a useful way to characterize a text or a genre and can be used to analyze lexico-grammatical features in a corpus. Thus, keyword analysis has great potential for application in linguistics and other relevant fields, including language teaching, forensic linguistics, stylistics, content analysis, and text retrieval (Scott 2008, O'Keeffe et al. 2007).

In general, keywords are computed using two word lists, one from the text or study corpus (SC) that one wants to investigate and the other from a normally larger, reference corpus (RC) that acts as a benchmark corpus or provides background data for keyword calculation. Since keyword calculation is performed basically on the basis of the comparison with the word list of the RC, its results are highly likely to be influenced by the RC chosen by a researcher. This makes us wonder what effects the size and composition of the RC have on keyword calculation and its results and what constitutes a good or bad RC. However, these important issues do not seem to have been sufficiently addressed in any previous studies, although the notion of keywords has attracted great attention and keyword analysis has been employed in many corpus-based language studies (e.g. Tribble 2000, Kemppanen 2004, Scott & Tribble. 2006, Seale et al. 2006, Goh & Lee 2008, Rayson 2008, Mahlberg 2009, McEnery 2009).

This paper is a quantitative study of the relationship between the RC and keyword calculation results. In particular, it will look into how various factors closely associated with the RC affect the keyword calculation of a given SC and its results. In analyzing keyword results and their differences, our discussion will be limited to the size of the keyword lists obtained from SCs in comparison with different RCs, thereby leaving the discussion of the composition of the keyword lists as a question for a future study.

The organization of this paper is as follows. We will first briefly review keyword calculation process and relevant previous studies. We will then investigate the roles of four major factors, including corpus size, genre, national variety, and diachrony, in keyword calculation, focusing on whether differences in any of these four major factors can cause a statistically significant difference in the number of keywords produced. We will also consider what the results of our analysis suggest about keyword analysis and its interpretation.

2. Background of the Study

2.1 Keyword Calculation

Corpus linguistics tools often contain a program or function which generates a keyword list of a text or corpus. For example, WordSmith Tools 5.0, one of the most widely used corpus linguistics tools, has a program called KeyWords that can be used to identify keywords in a text. This program generates the keywords of a text or corpus by comparing two word lists, one from the SC and the other from the RC and by calculating the keyness value of each word in the SC.

More specifically, to calculate the keyness value of a word, the KeyWords program computes the frequencies of the word in the SC and the RC, along with the number of running words in each of the two corpora, cross-tabulates the obtained numbers, and calculates the log-likelihood or Chi-Square of the word (Scott 2008). The result is a list of keywords, whose frequencies are unusually high (i.e. positive keywords) or unusually low (i.e. negative keywords). The figure below is a screen shot showing a keyword list from one of the twelve short stories (entitled *A Scandal in Bohemia*) in *The Adventures of Sherlock Holmes*. The RC used is the BNC (British National Corpus):

Figure 1. Keywords of *A Scandal in Bohemia*

adventure.1.kws

File Edit View Compute Settings Window Help

N	Key word	Freq.	%	RC. Freq.	RC. %	Keyness	P
1	HOLMES	48	0.56	866		526.26	0.0000000000
2	I	260	3.02	733,635	0.71	361.52	0.0000000000
3	BRIONY	11	0.13	7		182.70	0.0000000000
4	ADLER	13	0.15	109		161.59	0.0000000000
5	MY	78	0.91	146,879	0.14	158.49	0.0000000000
6	MAJESTY	16	0.19	536		156.00	0.0000000000
7	PHOTOGRAPH	21	0.24	2,599		150.64	0.0000000000
8	IRENE	13	0.15	508		122.80	0.0000000000
9	SHERLOCK	11	0.13	203		120.07	0.0000000000
10	HIS	105	1.22	410,351	0.40	95.23	0.0000000000

The figure above shows each keyword, along with its frequency and percentage in the SC, its frequency and percentage in the RC, its keyness, and p-value (which indicates the probability of being wrong in claiming a relationship). Note that the keyness value of *Holmes* is bigger than that of *I*, *my*, or *his*, although its raw frequency is much lower than the other keywords. Note also that although the definite article *the* is usually the most frequent word in a text, it is very unlikely to turn out to be a keyword in most texts. This is mainly because the keyness of a word is determined on the basis of the statistical significance of its frequency. Thus, a word will be included in the keyword list if its frequency is unusually high or unusually low in comparison with the frequency that would be expected on the basis of the word list from the RC.

2.2 Previous Studies

As is mentioned above, the calculation of keywords is based on the comparison with the RC, and therefore, the results of keyword calculation for a given text (i.e., the number of keywords and their composition) is highly likely to vary to a certain degree, depending on the RC that the researcher chooses. This possibility of variation in keyword results caused by using different RCs seems to be well recognized by corpus linguists (cf. Berber-Sardinha 2000, 2004; Scott 2009). Despite the great attention to the notion of keywords and a great many relevant studies employing this notion, however, the question of how the corpus size and composition of the RC affect keyword calculation and its results has not been sufficiently addressed in the literature. In fact, the only expectation or requirement for an appropriate RC appears to be that the RC should be larger in size than the SC, although even this requirement needs to be further verified and explained in more detail.

Thus, in this connection, Tribble (1999: 171) claims that the size of the RC is relatively unimportant. McEnery et al. (2006: 308-311) are in the same position as Tribble (1999) about the size of the RC. To verify Tribble's claim, they carry out a simple test involving the top ten positive keywords and top ten negative keywords in the two different keyword lists extracted from an American conversation on the basis of two different reference word lists from the one-million-word FLOB Corpus and the 100-million-word BNC. They find that nine out of the top ten positive keywords

and all the top ten negative keywords in the two keyword lists, respectively, are the same, in spite of the different sizes of the two RCs. McEnery et al. (2006) regard the result of this simple test as evidence to show the unimportance of the size of an RC in making a keyword list.

On the other hand, Berber-Sardinha (2000, 2004) provides a more serious and detailed discussion of the RC and its size. In particular, he finds, in a comparison of five SCs with RCs of various sizes, that an RC about five times the size of the SC yields a larger number of keywords than a smaller one. This finding above all means that an RC that is less than five times the size of the SC may not be reliable. Thus, although a larger RC is not always better than a smaller one (because using an extremely large corpus does not bring about a significant change in the number of keywords produced), he argues that the size of the RC matters in keyword calculation.

Scott (2009) is another study which attempts to address the issue of what constitutes a good or bad RC. As for the size of the RC and its role in keyword calculation, he asserts that changing the size of the RC does not cause a significant difference in the quality of keyword results, while genre-specific RCs identify rather different keywords. He also suggests that even an obviously absurd RC cannot be considered a bad RC because keywords identified by such an RC can still be plausible indicators of aboutness.

While insightful in certain respects, these studies have some important limitations. Above all, they were mostly concerned with the corpus size of the RC as a factor influencing keyword results, even though there are other important factors, especially those related to the composition of the RC, which are also highly likely to affect the results of keyword calculation. Although unlike other studies, Scott (2009) looked into how the genre difference of RCs affects keyword results on the basis of popularity and precision,[1] he fails to show in a more objective way whether 'different' keyword results generated by genre-specific RCs can be considered really different. Note that other aspects of the RC composition, such as diachronic and varietal differences, can also exert influence over keyword results. No

[1] Popularity, an indicator of quality or usefulness of keywords, is defined as presence of each keyword in at least 20 of the 22 genre-specific RC sets, while precision is computed following Oakes (1998: 176) and indicates "the proportion retrieved items that are in fact relevant (the number of relevant items obtained divided by the total number of retrieved items)".

previous study, however, seems to have investigated the possible effects that such factors have in keyword calculation.

Another limitation shared by all previous studies concerns the nature of the texts used as SCs in the analysis of the RC's role in keyword calculation. In particular, most SC texts used in their analyses are incomplete extracts from larger texts.[2] For example, four of the five SCs used in Berber-Sardinha (2000) are from the Brown Corpus, each of whose 500 text samples is an approximately 2,000 word long text fragment. Note that using short fragments instead of whole texts can skew the results of analysis because shorter texts allow less room for the repetition of words and phrases, thereby affecting word frequency. Furthermore, although it can also be used successfully with segments of texts, keyword analysis, other things being equal, can clearly have more meaning when it is used for a whole text or a set of whole texts which have continuity and unity in content.

3. Methodology

3.1 Corpus Data

The main goal of this paper is to investigate the effects that the size and composition of the RC can have on the results of keyword calculation. The texts or corpus data that we have analyzed to attain this goal were selected as follows. Above all, the texts to be used as the SCs for comparison with RCs of various kinds are two different series of short stories by Sir Arthur Conan Doyle, *The Adventures of Sherlock Holmes* and *The Return of Sherlock Holmes*. The two series contain 12 and 13 short stories (A01-A12, R01-R13), respectively, and the table below shows the basic statistics of the texts selected. Note that all the texts that we have chosen are whole texts, a selection considering the problem of using short fragments of larger texts in previous studies, which could skew or bias the test results.

---

[2] Note also that all previous studies have used only a small number of texts or corpora as SCs in their analyses. Thus, only five SCs and two SCs are used in Berber-Sardinha (2000) and Scott (2009), respectively.

Table 1. Study corpora

SC	token	type	SC	token	Type
A01	8,608	1,651	R01	8,747	1,726
A02	9,197	1,635	R02	9,304	1,596
A03	7,021	1,369	R03	9,702	1,585
A04	9,686	1,626	R04	7,885	1,492
A05	7,365	1,522	R05	11,511	1,873
A06	9,288	1,687	R06	8,181	1,568
A07	7,878	1,507	R07	6,774	1,431
A08	9,893	1,714	R08	8,384	1,468
A09	8,347	1,531	R09	6,507	1,208
A10	8,167	1,531	R10	8,995	1,561
A11	9,744	1,536	R11	8,079	1,541
A12	10,004	1,668	R12	9,227	1,537
			R13	9,735	1,638
A01-12	105,198	6,008	R01-13	113,031	6,107

On the other hand, we have used a set of different corpora or sub-corpora as the RCs for keyword calculation. In particular, the selection of the RCs has been made so that five major factors or variables, representing the size and composition of the RC, as given in the table below, can be properly considered in our analysis.[3]

[3] The BNC (British National Corpus) is a 100 million word collection of spoken and written British English (about 10% and 90%, respectively), while the ICE-GB (British Component of the International Corpus of English) is a one million word corpus of spoken and written British English (about 60% and 40%, respectively). The Brown Corpus (American English of the 1960s) forms the so-called Brown Family of English corpora together with three other comparative English corpora: the Frown Corpus (American English of the 1990s), the LOB Corpus (British English of the 1960s), and the FLOB Corpus (British English of the 1990s). Each of the four Brown Family corpora contains about one million words sampled from 15 categories. The three category groups of the Frown Corpus (i.e., ABC, J, and KLMNPR) represent Press (i.e., News), Learned (i.e., Academic Prose), and Fiction, respectively. These three category groups, along with Conversation, form four main genres or registers (cf. Biber et al. 1999).

Table 2. Variables for keyword calculation and RCs

Factors	RCs
Size	BNC vs. ICE-GB
Spoken vs. Written	ICE-GB: Spoken vs. Written
Major Registers	Frown: ABC, J, KLMNPR
National Varieties	Brown/Frown vs. LOB/FLOB
Diachrony	Brown/LOB vs. Frown/FLOB

The basic statistic information about the corpora used as the RCs is given in the following table:[4]

Table 3. Reference corpora

RC		Token	Type
BNC		96,986,707	361,660
ICE-GB	Written	423,702	19,516
	Spoken	637,562	16,686
	Total	1,061,264	26,159
Frown	ABC	177,178	14,839
	J	160,938	11,558
	KLMNPR	253,342	14,340
	Frown-All	1,003,051	35,576
Brown		1,014,312	32,583
FLOB		1,002,695	35,189
LOB		1,013,768	30,861

[4] The statistics given in the table are based on the figures officially published, while the number of the types used in each (sub)corpus has been obtained through lemmatization using WordSmith Tools 5.0 and the lemma list supplied by Yasumasa Someya and downloaded at http://www.lexically.net/downloads/e_lemma.zip. On the other hand, the number of the tokens used in each subgroup of the Frown Corpus is the result of combining the (officially given) numbers of the tokens of all subcategories in the subgroup. Note that Frown-All indicates all the tokens and types used in the Frown Corpus, not the total of the three subgroups.

3.2 Data Analysis

The software that we have used for keyword analysis in this study is WordSmith Tools 5.0 by Mike Scott, which is currently most widely used for corpus data analysis among (corpus) linguists and language teachers. Using this software package, consisting of three main programs, Concord, KeyWords and WordList, we produced the word lists of all the SCs and RCs and performed keyword calculation.

The software used for necessary statistical tests is R and SPSS 15.0 (data mining, statistical software).[5] The statistical computing function of R was used to determine the statistical significance of the difference(s) shown in each set of keyword results that were obtained through the comparison with the relevant RCs, while SPSS was used mainly for verifying the results of the statistical tests performed by R.

The general procedure for corpus data analysis that we have followed in this study is as follows: First, a set of SCs have been compiled from the texts of two series of Sherlock Holmes short stories (*The Adventures of Sherlock Holmes* and *The Return of Sherlock Holmes*), while different English (sub)corpora were selected and regrouped to be used as RCs. Second, the word list of each SC and RC has been generated along with relevant statistics, and all words in each word list have been lemmatized for a more precise analysis and comparison. Third, using each RC as background data for keyword calculation, a keyword list was produced from each of the 25 SCs. The settings for keyword calculation of the KeyWords program were as follows: minimum frequency = 3; maximum keywords = 500; negative keywords to be excluded; p value = 0.000001; statistical test for keyness calculation = log likelihood (Dunning 1993). Note that negative keywords, whose frequencies are statistically significantly low, were excluded. That is, only positive keywords were considered in the analysis of this study. Fourth, all the keyword lists obtained in comparison with the RCs of each factor group have been compared and tested for statistical significance of the given mean difference to determine whether the relevant factor brings about a statistically significant difference in the keyword results.

---

[5] R is a programming language and software environment for statistical computing and graphics. The R language is currently considered a de facto standard among statisticians for developing statistical software, and is widely used for statistical software development and data analysis (http://en.wikipedia.org/wiki/R_(programming_language) and http://www.r-project.org/).

Finally, all the results and findings have been examined to draw a conclusion about the research question and discuss their implications.

4. Results and Discussion

4.1 Corpus Size

What results are obtained if keyword calculations are performed using RCs of different sizes? Although this question has already been dealt with in some previous studies such as Berber-Sardinha (2000, 2004), as reviewed in Section 2.2, let us look at this question again and check the validity of their conclusion for a warm-up analysis. For this purpose, keyword calculations have been made for the 25 SCs using the ICE-GB and the BNC as two RCs of different sizes. The results of the keyword calculations are given in the following table.

Table 4. Keyword results (corpus size: ICE-GB vs. BNC)

Variable	N	M	SD
ICE-GB	25	65.64	11.99958
BNC	25	62.64	11.34637

As we can see in the table above, the mean numbers of the keywords obtained from the 25 SCs using two RCs of difference sizes are 65.64 and 62.64, respectively. With these mean numbers of keywords given, can we reasonably say that the size of the RCs is an important factor which brings about a statistically significant difference in keyword results? In order to answer this question, we performed a paired sample t-test. The test results, as given in the table below, indicate that the difference between the two sets of keyword results is statistically not significant ($p > 0.05$), thereby confirming the conclusion of previous studies:

Table 5. Paired sample T-test (size: ICE-GB vs. BNC)

variable	M	SD	t	p
ICE-GB - BNC	0.25	5.651087	0.9083	0.3683

4.2 Genre

Another factor which is highly likely to influence the results of keyword calculation is the genre or register difference of the RCs. What differences in keyword results do we obtain if genre difference is selected as a variable and keyword calculation is made in comparison with RCs of different genres? Let us first consider the case in which spoken texts and written texts are used as the RCs. The following table shows the keyword results obtained when the spoken and written parts of the ICE-GB are used as the RCs:

Table 6. Keyword results (genre: written vs. spoken)

RC	N	M	SD
Written (ICE-GB)	25	66.08	10.82790
Spoken (ICE-GB)	25	70.44	14.35584

In the table above, we can see that more keywords are obtained when spoken English texts are used as the RC. What do these keyword results tell us about the significance that the genre variation of spoken and written RCs has as a factor in keyword calculation?

Table 7. Paired sample T-test (genre: written vs. spoken)

Variable	M	SD	t	P
Written - Spoken	-0.2916667	12.93287	-1.2124	0.2318

As we can see in the table above, the difference in keyword results caused by this genre variation turned out to be statistically not significant ($p > 0.05$). This means that the genre difference of spoken and written RCs is not an important factor in keyword calculation. This result, considering the degree of difference between spoken and written language that we often expect or assume, is a little surprising.

Since the spoken and written genre difference, contrary to our expectation, turned out to be unimportant in keyword calculation, what would be the importance of other genre differences that the RCs show? In order to answer this question, we prepared three different RCs that represent three major genres or registers of written English,

that is, News, Academic Prose, and Fiction. These three RCs are from three different category groups of texts in the Frown Corpus: Press Categories A, B, and C for News, Learned Category J for Academic Prose, and Fiction Categories K, L, M, N, P, and R for Fiction. The results of keyword calculation using these three RCs are summarized in the following table:

Table 8. Keyword results (genre: news vs. prose vs. fiction)

RC	N	M	SD
Frown-ABC	25	90.76	14.641
Frown-J	25	124.56	19.292
Frown-KLMNPR	25	50.76	9.812

From the table above, we can see that the mean differences in keyword results are quite large, and this makes us expect the three-way genre difference of written English to turn out to be an important factor in keyword calculation. In order to confirm our expectation, a one-way repeated measures ANOVA test was conducted, as given in the following table:

Table 9. One-way repeated measures ANOVA, RMD (genre: news vs. academic vs. fiction)

Source	SS	df	MS	F	P
RC (A)	68241	2	34120	557.901	.000***
SC (B)	13553	24	565		
A x B	2834	48	59		
Total	84628	74			

The ANOVA test result shows ($p < 0.001$) that unlike the case of the spoken and written genre difference, there are significant differences in keyword results caused by the three-way genre difference of the RCs.

Note that more than two different genres are represented by the RCs. Thus, in order to see which two means are significantly different, we conducted a Tukey's HSD test, as follows.[6]

[6] Tukey's test, also known as the Tukey range test, Tukey method, Tukey's honest significance test, Tukey's HSD (Honestly Significant Difference) test, or the Tukey–Kramer method, is a

Table 10. Tukey's HSD test (genre: news vs. academic vs. fiction)[7]

RC		J	
		Frown-J	Frown-KLMNPR
I	Frown-ABC	-33.8**	40.0**
	Frown-J		73.8**

The results of Tukey's HSD test indicate that the two means of every pair are significantly different. This means that genre difference is important in keyword calculation when we choose written English texts as RCs. Note that the genre difference between spoken and written RCs, unlike the three-way genre difference of written RCs here, turned out to be unimportant. Thus, the significance of genre factor in keyword calculation should be stated with the specific genres involved clearly indicated.

4.3 National Varieties: American vs. British English

Over the past 400 years the English language used in the United States and that used in the United Kingdom have diverged in some ways, leading to the two major national dialects of English, generally referred to as American English and British English. Although these two varieties of English are very similar, there are still some differences in many aspects of language and the greatest difference is in their vocabularies including idioms and slangs (Swan 2005, Algeo 2006). Since keyword calculation is made on the basis of the lexical comparison of the words used in the SC and the RC, it would be interesting to see what effects varietal difference, especially between American English and British English, can have on keyword calculation and its results.

For this part of study, a keyword list for each of the 25 Sherlock Holmes stories as the SCs was computed again, using two pairs of American and British English corpora as the RCs: the Frown Corpus and the FLOB Corpus (for American and

---

single-step multiple comparison procedure and statistical test generally used in conjunction with an ANOVA to find which means are significantly different from one another. It compares all possible pairs of means on the basis of a studentized range distribution q which is similar to the distribution of t from the t-test (http://en.wikipedia.org/wiki/Tukey's_range_test).

[7] Numbers = mean of I – mean of J, **$p < 0.01$.

British English of the 1990s) and the Brown Corpus and the LOB Corpus (for American and British English of the 1960s). The following table shows the basic statistics of the keyword calculation results:

Table 11. Keyword results (varietal difference: AmE vs. BrE)

RC	N	M	SD
Frown	25	88.16	15.76568
FLOB	25	80.48	14.36175
Brown	25	77.96	14.64434
LOB	25	77.32	13.34703

From the table, we can easily see that there is a noticeable difference in the mean number of keywords between Frown and FLOB, whereas Brown and LOB are more similar than different in the mean number of keywords. What then does the mean difference, especially between Frown and FLOB, mean statistically? To answer the question, we performed a paired sample t-test for the two pairs of keyword calculation results, as given in the following table:

Table 12. Paired sample T-test (varietal differences: AmE vs. BrE)

Variable	M	SD	t	p
Frown - FLOB	0.08333333	4.959985	1.8006	0.07811
Brown - LOB	0.1666667	3.643537	0.1615	0.8724

As can be seen in the above table, the difference in mean caused by the varietal difference of American and British English turned out to be statistically not significant in either case ($p > 0.05$). Note, however, that the effect of varietal difference on keyword results is much greater when using Frown and FLOB (i.e., English varieties of the 1990s) as the RCs than when using Brown and LOB (i.e., English varieties of the 1960s), getting very close to a statistically significant level. Since keyword calculation is based on lexical comparison, this finding strongly suggests that the varietal difference in vocabulary has become greater between American and British English, although the two national varieties in general have been growing closer together since the beginning of the twentieth century (Algeo 2001: xix).

4.4 Diachrony

As is well known, language changes over time, leading to increasing differences between different periods of the same language. Diachronic or historical corpora (e.g., the Helsinki Diachronic Corpus of English Texts), which contain texts from the same language gathered from different time periods, are used mainly for the study of such diachronic changes. Although the Brown family corpora are generally not considered diachronic corpora, they can be used for a similar purpose, especially for the study of language change in progress, because they are designed for comparing language varieties not only synchronically but also diachronically.

Thus, using the two pairs of matching English corpora of the Brown family (i.e., Frown and Brown for American English and FLOB and LOB for British English), let us now analyze the possible effects of diachrony on the results of keyword calculation. Note that the results of keyword calculation for the 25 SCs are the same as those given in Table 11 above. This is because the RCs used for this part of study are the same as those used for analyzing the effects of varietal difference. In Table 11, we can see that the difference in the mean number of keywords between the two periods is greater in American English (88.16 - 77.96 = 10.20) than in British English (80.48 - 77.32 = 3.16).

What is the statistical significance of such mean differences? In other words, what are the effects that diachronic difference can have on keyword calculation when using texts from two different time periods as the RCs? In order to answer these questions, we performed a paired sample t-test for each case of diachrony. The test results are given in the following table:

Table 13. Paired sample T-test (diachrony: 1990s vs. 1960s)

Variable	M	SD	t	P
Frown - Brown	0.1666667	5.813527	2.3701	0.02187*
FLOB - LOB	0.25	5.391217	0.8059	0.4243

The statistical tests, as we can see in the table above, show that the mean difference is statistically significant only in the case of American English corpora ($p < 0.05$). This means that only in American English the diachronic change over the three decades exerted significant influence over the results of keyword calculation.

Since the two cases of diachrony resulted in conflicting test results, what should we say about the effects of diachrony on keyword calculation? The results of our analysis so far seem to allow us to say that diachrony is an important factor in keyword calculation, although the content of the diachronic difference involved needs to be clearly specified when we make such a statement.

5. Conclusion

The main goal of this paper was to explain what factors of the RCs influence the results of keyword calculation in a significant way. To achieve this goal, we have examined the possible effects that the size and composition of the RCs can have on keyword results. In particular, our analyses have shown that among the four major factors (i.e., corpus size, genre, varietal difference, and diachrony) contributing to different compositions of the RCs, only genre and diachrony can bring about statistically significant differences in the number of the keywords generated.

There are some important points to make about the results of our analyses in this paper. Above all, although genre and diachrony turned out to be important factors in keyword calculation, it needs to be more precisely stated in what specific cases they can exert significant influence over keyword results. This is because genre and diachrony, as we have seen in the two specific cases (i.e., the genre difference between spoken and written English and the diachronic change in British English), sometimes do not influence keyword results to a statistically significant extent.

Another point to note is that different keyword results made possible by using RCs of different compositions do not always have to be interpreted in connection with the question of what constitutes a good or bad RC or which RCs are better than others. Keyword analysis in general is very robust and keywords identified even by an obviously absurd RC are very likely to be plausible indicators of aboutness (Scott 2009). Thus, varying results of keyword calculation caused by varying the RC can be understood as arguing for the diversity of the aboutness of a text rather than different qualities of keyword lists or RCs.

Finally, this paper has one important limitation, especially in its scope, and this suggests directions for further research. That is, in analyzing the effects of the size and composition of the RCs on keyword computation, we have limited our study,

mainly for lack of space, to a quantitative analysis of keyword results. Since the composition of the keyword lists obtained with RCs of various kinds is equally important, one of the questions for further study will be how the composition of keyword lists changes depending on the use of different RCs.

References

Algeo, J. 2001. *The Cambridge History of the English Language*, Vol. VI, *English in North America*. Cambridge: Cambridge University Press.

Algeo, J. 2006. *British or American English?* Cambridge: Cambridge University Press.

Berber-Sardinha, T. 2000. Comparing corpora with WordSmith Tools: How large must the reference corpus be? *Proceedings of the Workshop on Comparing Corpora* 9, 7-13.

Berber-Sardinha, T. 2004. *Lingüuística de Corpus*. Brazil: Manole.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19.1, 61-74.

Goh, G-Y. and Lee, S-W. 2008. A Corpus-based analysis of the language of English news in Korea. *The Journal of Studies in Language* 23.4, 601-619.

Kemppanen, H. 2004. Keywords and ideology in translated history texts: A corpus-based analysis. *Across Languages and Cultures* 5.1, 89-106.

Mahlberg, M. 2009. Patterns in news stories. In L. Lombardo (ed.) *Using Corpora to Learn about Language and Discourse*. Bern: Peter Lang. pp. 99-132.

McEnery, T. 2009. Keywords and moral panics: Mary Whitehouse and media censorship. In D. Archer (ed.) *What's in Word-list? Investigating Word Frequency and Keyword Extraction*. Oxford: Ashgate.

McEnery, T., R. Xiao, and Y. Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

O'Keeffe, A., M. McCarthy, and R. Carter. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Rayson, P. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics* 13.4, 519-549.

Scott, M. 2008. *Oxford Wordsmith Tools 5.0*. Liverpool: Lexical Analysis Software.

Scott, M. 2009. In search of a bad reference corpus. In D. Archer (ed.) *What's in Word-list? Investigating Word Frequency and Keyword Extraction*. Oxford: Ashgate.

Scott, M. and C. Tribble. 2006. *Textual Patterns: Keyword and Corpus Analysis in Language*

Education. Amsterdam: John Benjamins.
Seale, C., Charteris-Black, and S. Ziebland. 2006. Gender, cancer experience and internet use: a comparative keyword analysis of interviews and online cancer support groups. *Social Science and Medicine* 62.10, 2577-2590.
Swan, M. 2005. *Practical English Usage*. Oxford: Oxford University Press.
Tribble, C. 1999. *Writing Difficult Texts*. Ph. D. thesis. Lancaster University.
Tribble, C. 2000. Genres, keywords, teaching: towards a pedagogic account of the language of project proposals. In L. Burnard and T. McEnery (eds.) *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the Third International Conference on Teaching and Language Corpora*. Frankfurt: Peter Lang. pp. 75-90.

Gwang-Yoon Goh
Dept. of English Language and Literature
Yonsei University
134 Shinchon-dong, Seodaemun-gu,
Seoul 120-749, Korea
E-mail: goh@yonsei.ac.kr

Received: 2011. 04. 05
Revised: 2011. 04. 21
Accepted: 2011. 04. 22